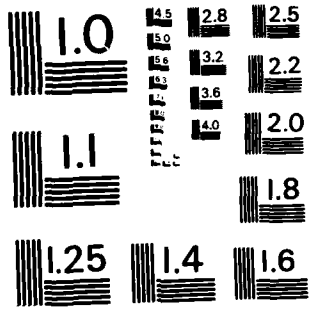
AD-A127 760 BALANCED SUBCLASSIFICATION IN OBSERVATIONAL STUDIES USING THE PROPENSITY..(U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER P R ROSENBAUM ET AL. 1/1

UNCLASSIFIED JAN 83 MRC-TSR-2468 DAAG29-80-C-0041 F/G 12/1 NL

END
DATE FILMED
5 83
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD A127760

MRC Technical Summary Report #2468

BALANCED SUBCLASSIFICATION IN OBSERVATIONAL STUDIES USING THE PROPENSITY SCORE: A CASE STUDY

Paul R. Rosenbaum
and
Donald B. Rubin

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

January 1983

(Received September 8, 1982)

DTIC FILE COPY

Approved for public release
Distribution unlimited

DTIC ELECTE MAY 06 1983 S D E

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Cancer Institute
9000 Rockville Pike
Bethesda, MD 20205

Educational Testing Service
Carter Road
Princeton, NJ 08541

83 05 06-135

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

BALANCED SUBCLASSIFICATION IN OBSERVATIONAL STUDIES USING THE PROPENSITY SCORE: A CASE STUDY

Paul R. Rosenbaum* and Donald B. Rubin**

Technical Summary Report #2468

January 1983

ABSTRACT

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Previous theoretical arguments have shown that subclassification on the scalar propensity score will balance all observed covariates. The procedure is illustrated in a large observational study of treatments for coronary artery disease. Five subclasses are constructed that balance 74 covariates. Balanced subclassification is combined with model-based adjustments to provide estimates of treatment effects within subpopulations. Two appendices address theoretical issues: (A) propensity scores from incomplete data, and (B) the effectiveness of subclassification on the propensity score.

AMS (MOS) Subject Classifications: 62F99; 62H99; 62P10; 62H17

Key Words: Observational studies; bias reduction; stratification; logistic models; log linear models; direct adjustment; balancing scores.

Work Unit Number 4 - Statistics and Probability

* Departments of Statistics and Human Oncology, University of Wisconsin-Madison.

** Departments of Statistics and Education, University of Chicago.

Sponsored in part by the United States Army under Contract No. DAAG29-80-C-0041, Grant P30-CA-14520 from the U.S. National Cancer Institute to the Wisconsin Clinical Cancer Center, and by the Wisconsin Alumni Research Foundation, the Educational Testing Service, the U.S. Health Resources Administration.

SIGNIFICANCE AND EXPLANATION

In observational studies, treatments are assigned to experimental units without the benefit of randomization. As a result, the units receiving the various treatments may not be comparable with respect to observable or unobservable characteristics. A detailed example that builds on previous theoretical arguments shows that it is possible to form a few groups or subclasses of units such that, within each subclass, units receiving different treatments are comparable with respect to the distribution of observable characteristics.

Accession For	
NTIS GRA&I	☒
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist | Avail and/or Special	
A	

DTIC COPY INSPECTED 8

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

BALANCED SUBCLASSIFICATION IN OBSERVATIONAL STUDIES USING THE PROPENSITY SCORE: A CASE STUDY

Paul R. Rosenbaum* and Donald B. Rubin**

1. Introduction: Subclassification; the Propensity Score; a Case Study

1.1. Adjustment by Subclassification In Observational Studies

In observational studies for causal effects, treatments are assigned to experimental units without the benefits of randomization. As a result, treatment groups may differ systematically with respect to relevant characteristics, and therefore not be directly comparable. One commonly used method of controlling for systematic differences involves grouping units into subclasses based on observed characteristics, and then directly comparing only treated and control units who fall in the same subclass.

Cochran (1968) presents an example in which the mortality rates of cigarette smokers, cigar/pipe smokers and nonsmokers are compared after subclassification on the covariate age. The age-adjusted estimates of the average mortality for each type of smoking were found by direct adjustment, that is, by combining the subclass-specific mortality rates using weights equal to the proportions of the population within the subclasses. Cochran (1968) shows that five subclasses are often sufficient to remove over 90% of the bias due to the subclassifying variable or covariate. However, as noted by Cochran (1965), as the number of covariates increases, the number of subclasses grows exponentially, so that even with only two categories per covariate, yielding 2^p subclasses with p covariates, some subclasses will contain no units, and many subclasses will contain either treated or control units but not both, making it impossible to form directly adjusted estimates for the entire population.

* Departments of Statistics and Human Oncology, University of Wisconsin-Madison.

** Departments of Statistics and Education, University of Chicago.

Sponsored in part by the United States Army under Contract No. DAAG29-80-C-0041, Grant P30-CA-14520 from the U. S. National Cancer Institute to the Wisconsin Clinical Cancer Center, and by the Wisconsin Alumni Research Foundation, the Educational Testing Service, the U. S. Health Resources Administration.

Fortunately, however, there exists a scalar function of the covariates, namely the propensity score, that summarizes the information required to balance the distribution of the covariates. Specifically, subclasses formed from the scalar propensity score will balance all p covariates. In fact, often five subclasses constructed from the propensity score will suffice to remove over 90% of the bias due to each of the covariates.

1.2. The Propensity Score in Observational Studies

In a study comparing two treatments, labeled 1 and 0, the propensity score, $e(\underset{\sim}{x})$, is the conditional probability that a unit with vector $\underset{\sim}{x}$ of observed covariates will be assigned to treatment 1, $e(\underset{\sim}{x}) = pr(z=1|\underset{\sim}{x})$, where $z = 1$ or 0 indicates the treatment assignment. Rosenbaum and Rubin (1983a, Theorem 1) show that subclassification on the population propensity score will balance $\underset{\sim}{x}$, in the sense that within subclasses that are homogeneous in $e(\underset{\sim}{x})$, the distribution of the observed covariates $\underset{\sim}{x}$ is the same for treated and control units. Formally, $\underset{\sim}{x}$ and z are conditionally independent given $e(\underset{\sim}{x})$, or in Dawid's (1979) notation:

$$\underset{\sim}{x} \perp\!\!\!\perp z | e(\underset{\sim}{x}) \quad . \tag{1}$$

The propensity score will balance $\underset{\sim}{x}$ whether or not $\underset{\sim}{x}$ includes all of the covariates used to assign treatments.

Under the assumption of strongly ignorable treatment assignment, defined by Rosenbaum and Rubin (1983a), appropriate adjustment for the propensity score alone will produce unbiased estimates of treatment effects. One way in which treatment assignment would be strongly ignorable is if, at each value of the observed covariates $\underset{\sim}{x}$, treatments are assigned randomly with positive probability to each treatment. A method for assessing the sensitivity of conclusions to violations of this assumption of strong ignorability has been described by Rosenbaum and Rubin (1983b) in a particular case; methods of testing strong ignorability have been reviewed by Rosenbaum (1982).

Subclassification on the propensity score is not the same as any of the several methods proposed by Miettinen (1976): the propensity score is not generally a "confounder" score. See Rosenbaum and Rubin (1983a, §3.3) for discussion.

1.3. A Case Study of Balanced Subclassification

In this paper, we illustrate balanced subclassification on the propensity score in an observational study of two treatments for coronary artery disease: coronary artery bypass surgery (z=1) and medical therapy (z=0). The vector of covariates $\underset{\sim}{x}$ contains 74 hemodynamic, angiographic, laboratory and exercise test results. The data analysis that follows is intended for purposes of illustration, and does not constitute a study of coronary bypass surgery.

The propensity score was estimated using a logit model (Cox 1970) for z,

$$\log \frac{e(\underset{\sim}{x})}{1-e(\underset{\sim}{x})} = \alpha + \underset{\sim}{\beta}^T \underset{\sim}{f}(\underset{\sim}{x})$$

where α and $\underset{\sim}{\beta}$ are parameters, and $\underset{\sim}{f}(\cdot)$ is a specified function.

2. Fitting the Propensity Score; Assessing the Balance Within Subclasses

2.1. The First Fit and Subclassification

Not all of the 74 covaraites and their interactions were included in the logit model for the 1515 patients in the study. Variables were selected for inclusion in the logit model using a stepwise procedure. A second stepwise selection added cross-products or interactions of those variables that were selected by the first stepwise procedure.

Based on Cochran's (1968) results and a new result in Appendix B of this paper, we may expect approximately a 90% reduction in bias for each of the 74 variables if we subclassify at the quintiles of the distribution of the population propensity score. Consequently, we subclassified at the quintiles of the distribution of the estimated propensity score based on this initial analysis, which we term the first model.

We now examine the balance achieved by this subclassification. Each of the 74 covariates was subjected to a two-way (2×5 = treatment x subclass) analysis of variance. Column 1 of Table 1 displays the F-ratios, that is, the squares of the usual two sample t-statistics for comparing the medical and surgical group means for each covariate prior to subclassification. Columns 2 and 3 display the F-ratios for the main effect of the treatment and the treatment-by-subclass interaction in the two-way analysis of

Table 1

F-TESTS OF BALANCE BEFORE AND AFTER STRATIFICATION

		Model #1		Model #2		Model #3		Final Model	
Variable	2-Sample F-statistic	Main Effect	Inter-action	Main Effect	Inter-action	Main Effect	Inter-action	Main Effect	Inter-action
AGE	4.4	.0	1.2	.0	1.0	.0	.9	0.0	0.7
PAINTYP	18.1	.0	.1	.0	1.8	.0	.6	0.0	0.7
CPAINSEV	6.8	.0	.6	.0	.6	.0	1.4	0.0	1.4
√CPAINFR	25.0	.1	.6	.1	.2	.1	1.8	0.2	0.8
CPAINTIM	5.3	.5	.2	.8	.1	1.1	1.4	1.0	0.9
TEMP	7.3	1.6	.8	.3	1.0	2.6	1.5	2.0	1.2
SEXACT	26.0	.2	.4	.0	.0	.2	.9	0.2	0.3
REST	10.9	2.2	.9	1.0	.6	1.9	.8	1.6	0.5
NITROGLY	11.6	1.5	.8	1.0	1.7	1.0	2.3	1.2	1.0
ADRENERG	6.8	.1	1.2	.2	.9	.1	1.1	0.1	1.2
CPAINCOR	38.4	.8	1.2	.0	.6	.3	.7	0.4	1.4
PREINFAR	9.0	4.6	1.7	.0	1.3	.0	2.2	0.1	2.9
HXCHF	6.8	.4	1.3	.0	.5	.0	1.2	0.0	0.8
CHF_SEV	7.3	.4	.9	.0	.6	.1	.6	0.1	0.4
DIG_CRX	4.4	7.1	1.0	.0	.5	.0	.1	0.0	0.2
BAB_CRX	23.0	.0	.7	.0	1.3	.2	.4	0.0	0.6
NITG_CRX	10.2	.7	1.0	.3	.9	.3	1.7	0.3	1.1
LNIT_CRX	31.4	.0	.4	.1	.8	.2	.7	0.1	2.2
SYSBP	4.8	4.0	.2	.3	.7	.2	.9	0.1	0.7
DIASBP	6.2	4.6	.0	.0	.4	.1	1.0	0.1	0.9
√FUNDUS	20.2	.0	1.9	.0	.8	.3	.5	0.2	1.3
NLPRECO	7.8	.5	.2	.3	.3	.3	1.7	0.5	0.9
ABNLPREC	10.2	.8	.2	.3	.3	.4	1.5	0.6	0.8
VENGALL	4.8	.1	.7	.0	.2	.1	.0	0.2	0.0
RFEMBRUI	6.8	4.0	.8	.0	1.9	.0	1.7	0.0	1.3
CMG	25.0	.3	2.9	.0	1.4	.2	.3	0.2	0.0
EKGPW	10.9	4.0	.4	3.3	.7	.3	1.0	0.2	0.3
EKGLAE	10.9	4.4	.3	3.3	.8	.2	.9	0.2	0.2
EKVIGCD	4.0	1.0	.4	.6	.2	.0	.8	0.0	1.3
EKGRAD	5.8	3.4	.9	.7	.2	.1	.1	0.1	0.1
STTW	8.4	.1	1.1	.1	1.5	.2	.9	0.3	0.5
STTPTC	13.0	.0	.4	.0	1.2	.1	.8	0.1	0.2
†TMSTGE	13.0	3.4	3.0	3.1	2.0	1.8	.3	2.1	0.4
†TMMAXHR	16.0	1.8	.5	.0	.8	.1	1.6	0.1	1.4

Table 1 (Continued)

†ADS	24.0	7.2	2.3	.0	1.2	.3	.7	0.3	0.1
PAS	16.0	3.5	1.2	.8	.2	1.1	.2	1.0	0.2
PAD	9.6	2.5	2.5	.7	.2	.8	.4	0.7	0.4
PAM	10.9	2.8	1.5	.6	.2	1.1	.3	0.7	0.2
PCWM	4.0	.0	.6	.2	.8	.2	.8	0.2	0.8
LVEDP	14.4	3.0	.1	.0	.4	.2	.5	0.1	0.4
LVCNL	7.8	4.8	.2	.6	2.6	.1	1.4	0.7	0.8
*LVCDAC	51.8	1.0	2.0	.1	1.0	.5	.4	0.4	0.9
LVCINF	14.4	7.0	1.8	.2	.2	.0	.5	0.1	0.4
LVCANT	9.6	1.2	1.3	.4	1.1	.8	.5	1.0	1.3
HYPOKIN	29.2	2.8	1.0	.4	.9	.4	.7	0.3	0.4
√MI	4.3	3.6	3.4	1.7	.7	1.1	.6	0.5	0.8
VVC	18.5	9.5	1.6	.0	1.3	.4	2.0	0.3	2.2
EDV	7.8	.2	1.5	.0	.3	.2	.7	0.4	0.5
ESV	15.2	.9	2.0	.0	.6	.2	.5	0.4	0.2
LVSV	5.8	1.7	.8	.0	.6	.0	.9	0.0	0.8
†EJFX	19.4	1.2	2.0	.1	1.1	.0	.3	0.0	0.3
ASYNSEPT	5.8	3.3	.7	1.2	.8	2.2	1.0	2.3	1.4
AKININF	13.0	3.0	1.4	2.2	.6	2.2	.4	3.6	2.0
AKINPOST	6.2	1.9	.6	1.3	.9	.4	.2	0.6	0.8
DYSKANT	8.4	1.8	.3	1.2	.6	.8	.7	0.9	0.9
DYSKPIC	16.0	4.2	1.6	2.5	.0	1.0	.2	1.1	0.4
MIDRCAST	5.8	1.6	.2	.6	.7	1.7	.7	1.6	0.8
PROXBEF	6.8	1.2	.4	.2	.7	.1	1.1	0.2	2.4
HXMI	4.8	1.0	1.3	.1	.3	.6	1.7	0.5	0.9
†TMPOS	14.4	1.4	1.1	.8	2.5	.5	.8	0.2	0.6
LAD_SIG	7.8	1.3	3.2	.1	5.1	.0	2.1	0.0	1.1
*LMC_SIG	22.1	1.2	1.2	.1	.8	.2	.4	0.3	0.2
LCA_SIG	6.2	4.2	1.2	.0	1.4	1.7	.1	1.0	1.2
EDVI	11.6	.2	1.4	.0	.4	.0	.5	0.2	0.3
ESVI	18.5	1.1	1.9	.0	.6	.2	.4	0.3	0.2
PROGPAIN	43.6	.3	.6	.0	.2	.1	1.1	0.1	1.4
STABPAIN	31.4	.0	.3	.1	.1	.0	1.5	0.0	1.0
TYPANG	18.5	.0	.1	.0	1.8	.0	.6	0.0	0.7
CPSEV	13.0	.0	.8	.4	1.2	.7	1.7	0.8	2.3
*NO1	10.9	.0	.4	.0	1.0	.0	.8	0.0	2.1
*NO2	10.9	.2	.2	.0	1.6	.0	2.6	0.0	2.4
LAD_DIST1	11.6	-	-	-	-	.1	2.4	0.4	1.4
LCA_DIST1	16.8	-	-	-	-	3.0	.6	0.0	1.2
RCA_DIST1	7.8	-	-	-	-	.5	.5	3.1	0.5

*Figures 3, 4 and 5 refer to variables (NO1,NO2), LVCDAC and LMC_SIG, respectively. See §2.2.
†Indicates a variable that was missing for at least 30% of patients. See §2.4 and Appendix A.
-Not included in early analyses.

variance. Although there has been a substantial reduction in most F-ratios as compared with column 1, several of the F-ratios are still quite large, possibly indicating that the propensity score is poorly estimated by the current model. Indeed, as a consequence of Theorem 1 of Rosenbaum and Rubin (1983a), each such F-test is an approximate test of the adequacy of the model for the propensity score; the test is only approximate primarily because the subclasses are not exactly homogeneous in the fitted propensity score.

2.2. Refinement of the Fitted Propensity Score; Balance Obtained in the Final Subclassification

Columns 4 through 7 describe subclasses based on a sequence of models. At each step, variables with large F-ratios that had previously been excluded from the model were added. For variables with large F-ratios that had previously been included in the model, cross-product terms were added. The F-ratios for the final model appear in columns 8 and 9, and are plotted in Figures 1 and 2. There is considerably greater balance within these final subclasses than would have been expected from randomized assignment to treatment within subclasses.

Figures 3, 4 and 5 display the balance within subclasses for three important covariates. Although the procedure used to form the subclasses may not be accessible to some nonstatisticians, the comparability of patients within subclasses can be examined with the simplest methods, such as the bar charts used here. For example, Figure 4 indicates some residual imbalance on the percent of patients with poor LV contraction, at least for patients in subclass 1, that is, in the subclass with the lowest estimated probabilities of surgery. This imbalance is less than would be expected from randomization within subclasses; see LVCDAC in Table 1. Nontheless, we would possibly want to adjust for this residual imbalance, perhaps using methods described in §3.3.

2.3. The Fitted Propensity Score; Overlap of Treated and Control Groups

Figure 6 contains boxplots (Tukey, 1977) of the final fitted propensity scores. By construction, most surgical patients have higher propensity scores, that is higher estimated probabilities of surgery, than most medical patients. There are a few surgical patients with higher estimated probabilities of surgery than any medical patient,

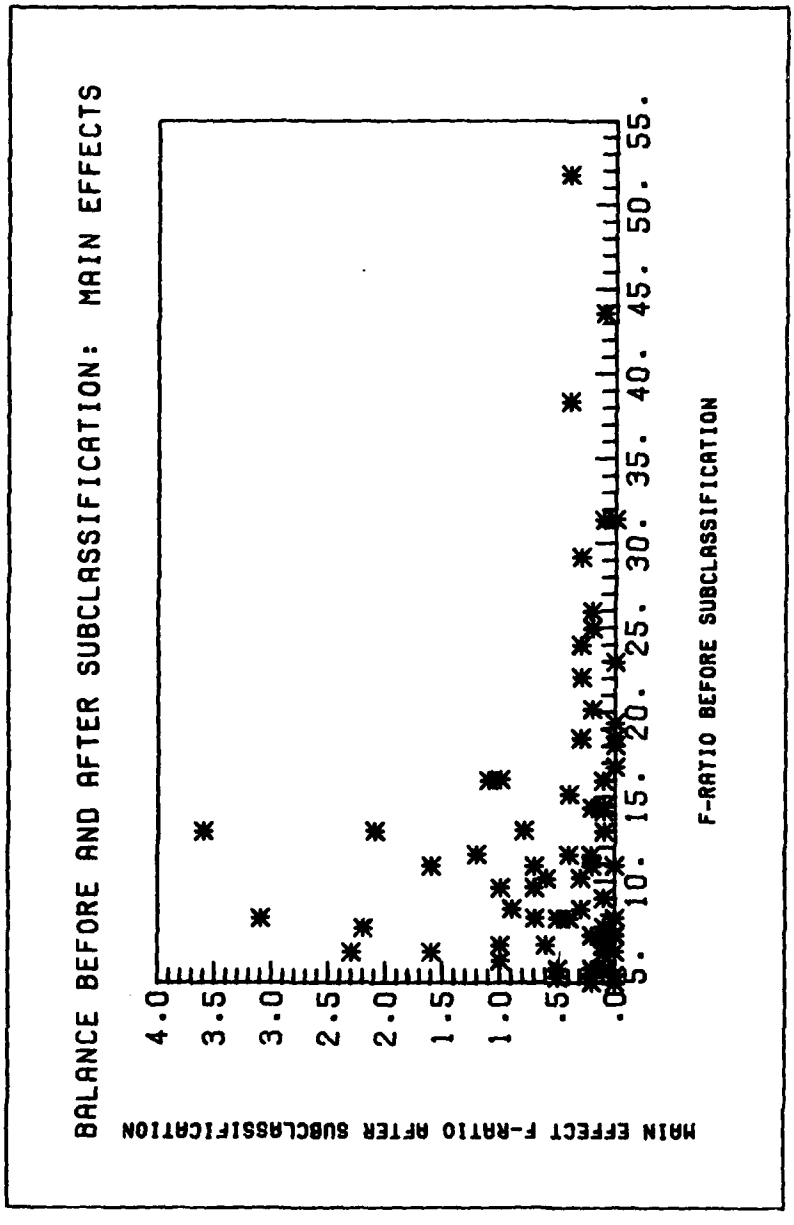

BALANCE BEFORE AND AFTER SUBCLASSIFICATION: MAIN EFFECTS
MAIN EFFECT F-RATIO AFTER SUBCLASSIFICATION
4.0
3.5
3.0
2.5
2.0
1.5
1.0
.5
.0
5.
10.
15.
20.
25.
30.
35.
40.
45.
50.
55.
F-RATIO BEFORE SUBCLASSIFICATION

Figure 1

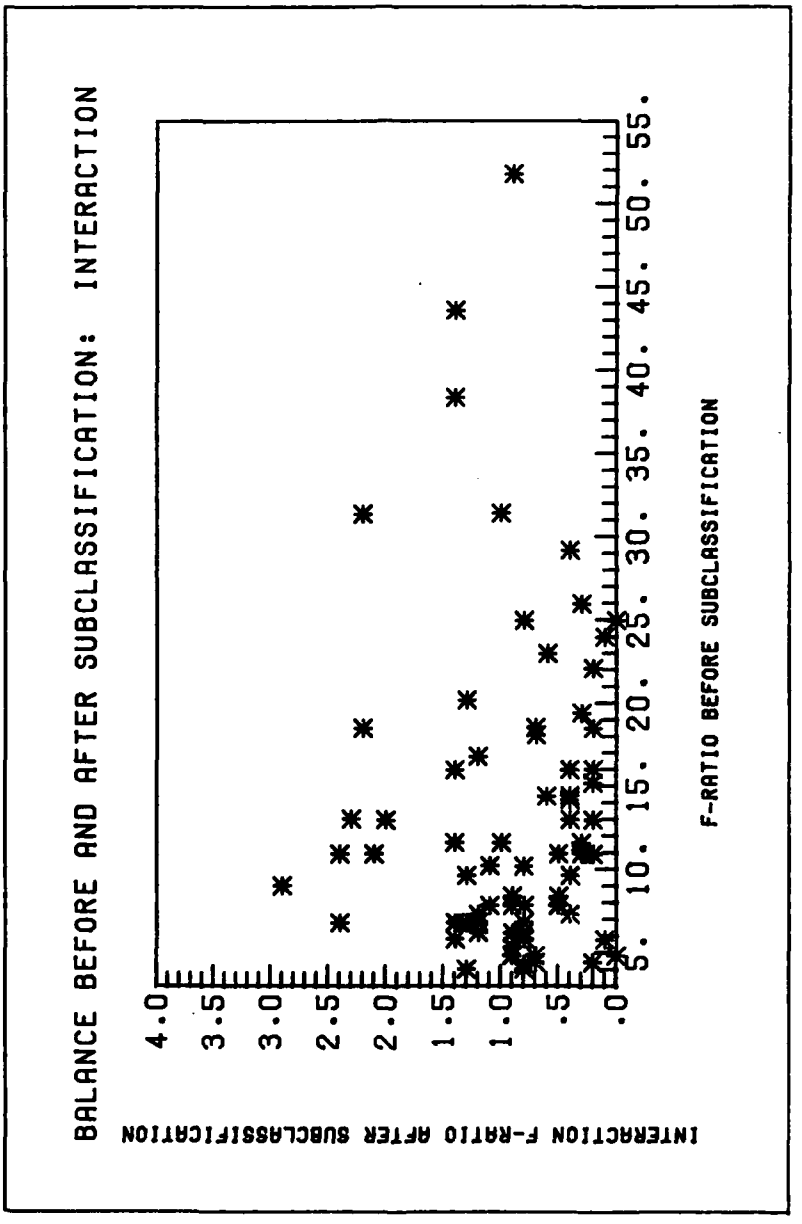

BALANCE BEFORE AND AFTER SUBCLASSIFICATION: INTERACTION
INTERACTION F-RATIO AFTER SUBCLASSIFICATION
F-RATIO BEFORE SUBCLASSIFICATION
4.0
3.5
3.0
2.5
2.0
1.5
1.0
.5
.0
5.
10.
15.
20.
25.
30.
35.
40.
45.
50.
55.

Figure 2

Figure 3

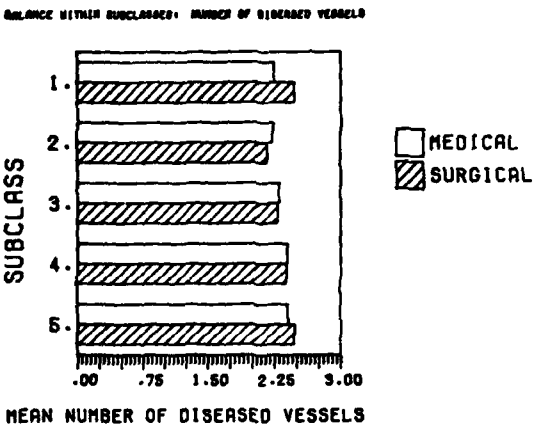


Figure 4

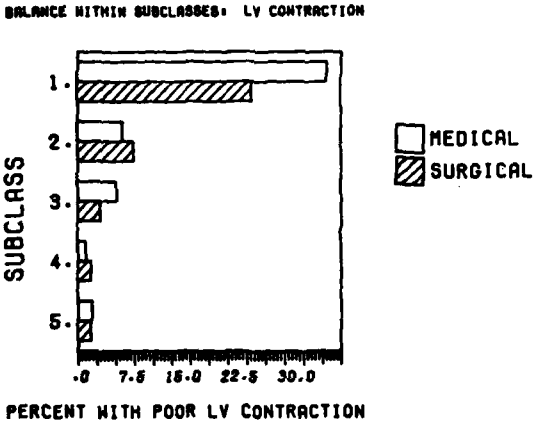


Figure 5

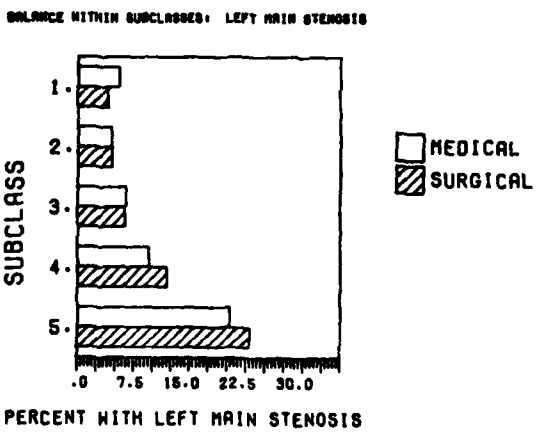
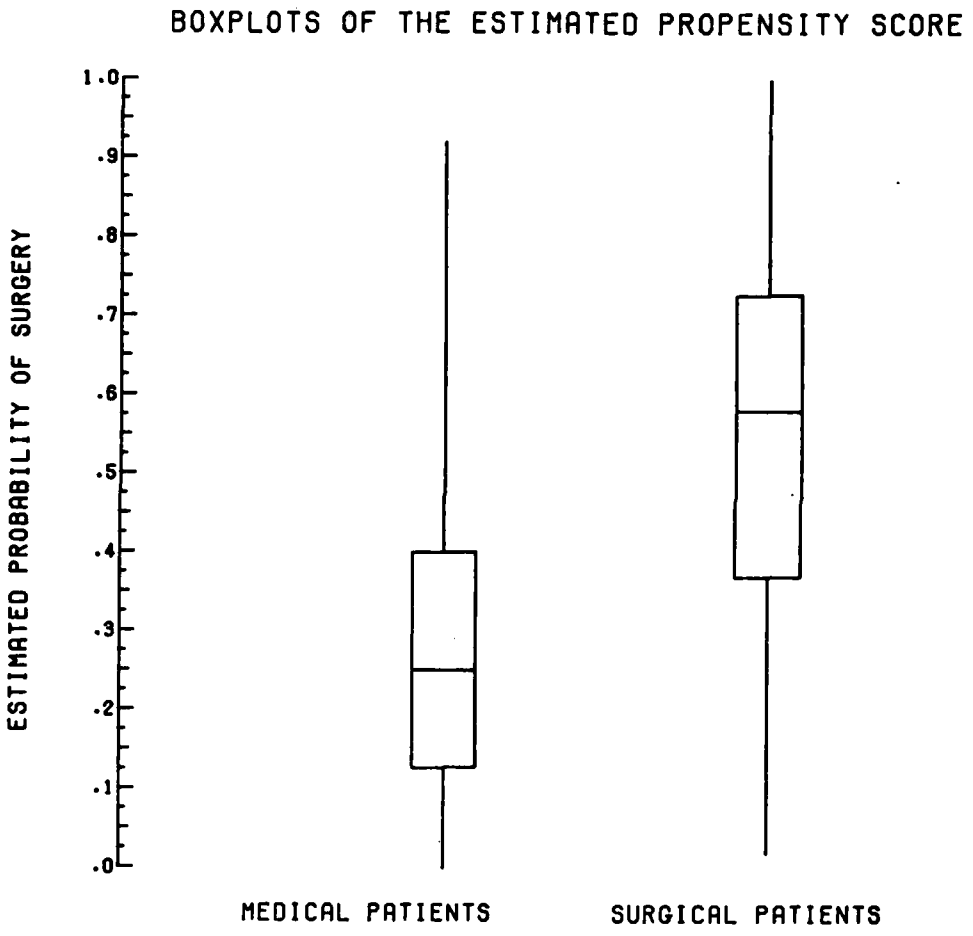


Figure 6


BOXPLOTS OF THE ESTIMATED PROPENSITY SCORE
ESTIMATED PROBABILITY OF SURGERY
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
.0
MEDICAL PATIENTS
SURGICAL PATIENTS

indicating a combination of covariate values not appearing in the medical group. For almost every medical patient, however, there is a surgical patient who is comparable in the sense of having a similar estimated probability of surgery.

2.4. Incomplete Covariate Information

The variables in Table 1 that are identified by a dagger (†) were not measured during the early years of the study, so that many patients are missing these covariate values. If the propensity score is defined as the conditional probability of assignment to treatment 1 given the observed covariate information and the pattern of missing data, then Appendix A shows that subclassification on the propensity score will balance both the observed data and the pattern of missing data. Essentially, we estimated the probabilities of surgical treatment separately for early and late patients, and then used these estimated probabilities as propensity scores. Subclassification on the corresponding population propensity scores can be expected to balance, within subclasses, each of the following: (a) the distribution of those covariates that are measured for both early and late patients, (b) the proportions of early and late patients, (c) the distribution of all covariates for the late patients. (For proof, see Corollary 1.1 of Appendix A.) Table 1 shows that the observed values of all covariates were indeed balanced by our procedure.

3. Results: Estimates of the Average Treatment Effect

3.1. Functional Improvement as the Response Variable; Placebo Effects

In this section, medicine and surgery are compared with respect to a particular response, namely functional improvement. Functional capacity is measured by the crude, four category (I = best, II, III, IV = worst) New York Heart Association classification, which measures a patient's ability to perform common tasks. The current study is confined to patients in classes II, III, or IV at the time of cardiac catherterization, i.e., patients who could improve. A patient is defined to have substantial improvement at 6 months after cardiac catherterization if he:

1. is alive, and
2. has not had a myocardial infarction, and

3. is in class I, or has improved by two classes

(i.e., IV to II);

otherwise, the patient is not substantially improved at six months.

It should be noted that there is substantial evidence that patients suffering from coronary artery disease respond to placebos; for a review of this evidence, see Benson and McCallie (1979). Part or all of the treatment effect may reflect differences in the placebo effects of the two treatments.

3.2. Subclass Specific Results; Direct Adjustment

The proportions improved in each subclass for medicine and surgery are displayed in Figure 7. In each subclass, the proportion improved under surgery exceeds the proportion improved under medical therapy.

Each subclass contains a total of 303 patients. Therefore, for medical therapy and surgery, the directly adjusted proportions improved, with subclass total weights, are simply the averages of the five subclass specific proportions. These adjusted proportions improved are .36 for medicine and .67 for surgery. Standard errors for the adjusted proportions, calculated following Mosteller and Tukey (1977, Chapter 11c), are .04 and .05, respectively. By Corollary 4.2 of Rosenbaum and Rubin (1983a), if treatment assignment is strongly ignorable, the difference between the surgical and medical adjusted proportions would be asymptotically unbiased for the average treatment effect if subclasses were exactly homogeneous in the estimated propensity score. The results of Cochran (1968) and Appendix B suggest that adjustment with the five subclasses used here will remove about 90% of the initial bias, providing treatment assignment is strongly ignorable. Similar considerations apply to the subpopulation specific results described in §3.3.

3.3. Adjustment and Estimation Within Subpopulations Defined by $\underset{\sim}{x}$

This section obtains adjusted estimates of the probabilities of substantial improvement at six months for subpopulations of patients defined by the number of diseased vessels (N) and the New York Heart Association functional class at the time of cardiac catherterization (F). To avoid an excessive number of subpopulations, the small but

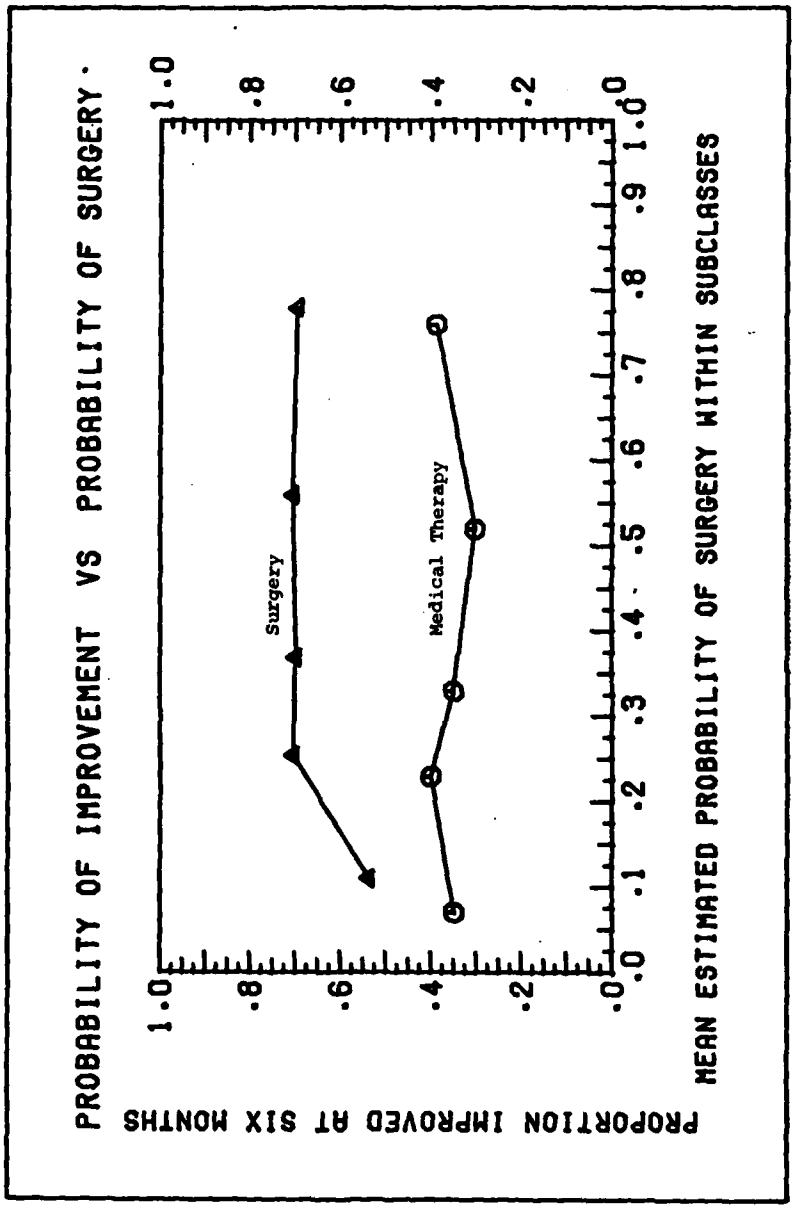

PROBABILITY OF IMPROVEMENT VS PROBABILITY OF SURGERY
PROPORTION IMPROVED AT SIX MONTHS
MEAN ESTIMATED PROBABILITY OF SURGERY WITHIN SUBCLASSES
Surgery
Medical Therapy
0.0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0

Figure 7

clinically important subset of patients with significant left main stenosis has been excluded.

In Table 2, patients are cross-classified according to the number of diseased vessels (N), initial functional class (F), treatment (Z), subclass (S), and condition at six months (I; improved = substantial improvement as defined in § 3.1). A loglinear model which fixed the IZN, IZF, ISN, SZ, SF, FN margins provides a good fit to this table (likelihood ratio chi square 122.5 on 120 degrees of freedom). (Here, IZN denotes the marginal table formed by summing the entries in the table over initial functional class F and subclass S, leaving a three way table.)

The directly adjusted estimates in Table 3 were calculated from the fitted counts using the NFS marginal table for weights. In other words, within each subpopulation defined by the number of diseased vessels (N) and the initial functional class (F), estimates of the probabilities of improvement are adjusted using subclass (S) total weights. For example, from Table 2, the weight applied to both medical and surgical estimated probabilities at N = ONE, F = II, subclass 1 is proportional to $9 + 8 + 1 + 0 = 18$.

The key observations from Table 3 are the following:

1. In all six subpopulations, the estimated probabilities of substantial improvement at six months are higher following surgery than following medical treatment (between 30% and 387% higher). The estimated probabilities differ least for one vessel disease, functional class IV, and differ most for three vessel disease, functional class III. As noted above, these differences may reflect differences in placebo effects of the two treatments (Benson and McCalie, 1979).

2. The definition of substantial improvement has resulted in lower estimated probabilities of improvement for class III patients than for class II and class IV patients.

3. The estimated probabilities of improvement under surgery vary less than the estimated probabilities of improvement under medicine.

Table 2. Counts Within Subpopulations

Number of Diseased Vessels	Initial Functional Class	Treatment	Subclass	Condition at 6 Months (I)	
N	F	Z	S	Improved	Not Improved
ONE	II	MEDICAL	1	9	8
			2	6	4
			3	2	4
			4	1	3
			5	0	0
		SURGICAL	1	1	0
			2	3	2
			3	1	1
			4	0	1
			5	1	0
	III	MEDICAL	1	4	9
			2	1	9
			3	1	4
			4	1	2
			5	0	1
		SURGICAL	1	0	0
			2	2	0
			3	1	2
			4	5	1
			5	3	1
	IV	MEDICAL	1	27	20
			2	15	15
			3	10	13
			4	5	6
			5	2	3
		SURGICAL	1	1	2
			2	10	3
			3	4	6
			4	8	6
			5	12	5
TWO	II	MEDICAL	1	3	4
			2	9	10
			3	6	9
			4	1	10
			5	0	1
		SURGICAL	1	3	0
			2	0	0
			3	6	4
			4	4	0
			5	2	2

Table 2 (continued)

	III	MEDICAL	1	2	4
			2	2	9
			3	4	12
			4	8	4
			5	1	2
		SURGICAL	1	0	0
			2	1	3
			3	4	1
			4	9	5
			5	11	5
	IV	MEDICAL	1	7	8
			2	12	15
			3	16	15
			4	6	11
			5	6	7
		SURGICAL	1	2	1
			2	5	1
			3	8	2
			4	25	7
			5	27	13
THREE	II	MEDICAL	1	5	25
			2	9	9
			3	2	8
			4	2	4
			5	0	2
		SURGICAL	1	3	0
			2	2	0
			3	4	1
			4	7	1
			5	8	4
	III	MEDICAL	1	5	24
			2	3	23
			3	4	14
			4	1	13
			5	1	5
		SURGICAL	1	2	5
			2	5	4
			3	7	3
			4	17	5
			5	14	6
	IV	MEDICAL	1	14	62
			2	21	39
			3	15	40
			4	8	29
			5	3	11
		SURGICAL	1	1	4
			2	13	7
			3	20	7
			4	22	15
			5	36	18

TABLE 3

Directly Adjusted Estimated Probabilities of Substantial Improvement

M=medical therapy
S=surgery

		Initial Functional Class		
		II	III	IV
Number of Diseased Vessels	One	M .469 S .708	M .277 S .629	M .487 S .635
	Two	M .404 S .780	M .221 S .706	M .413 S .714
	Three	M .248 S .709	M .133 S .649	M .278 S .657

4. Sensitivity of Estimates to the Assumption of Strongly Ignorable Treatment Assignment

The estimates in Section 3 are approximately unbiased under the assumption of strongly ignorable treatment assignment. Rosenbaum and Rubin (1983b) develop and apply to the current example a method for assessing the sensitivity of these estimates to a particular violation of strong ignorability. They assume that treatment assignment is not strongly ignorable given the observed covariates $\underset{\sim}{x}$, but is strongly ignorable given $(\underset{\sim}{x},u)$, where u is an unobserved binary covariate. The estimate of the average treatment effect is recomputed under various assumptions about u. A related Bayesian approach is developed by Rubin (1978).

5. Conclusions: The Propensity Score and Multivariate Subclassification

With just five subclasses formed from an estimated scalar propensity score, we have substantially reduced the bias in 74 covariates simultaneously. Although the process of estimating the propensity score for use in balanced subclassification does require some care, the comparability of treated and control patients within each of the final subclasses can be verified using the simplest statistical methods, and therefore results based on balanced subclassification can be persuasive even to audiences with limited statistical training. The same subclasses can also be used to estimate treatment effects within subpopulations defined by the covariates $\underset{\sim}{x}$. Moreover, balanced subclassification may be combined with model based adjustments to provide improved estimates of treatment effects within subpopulations.

Appendix A: Balancing Properties of the Propensity Score with Incomplete Data

In §2.4, we noted that several covariates were missing for a large number of patients. Let $\underset{\sim}{x}^*$ be a p-coordinate vector, where the j^{th} coordinate of $\underset{\sim}{x}^*$ is a covariate value if the j^{th} covariate was observed, and is an asterisk (*) if the j^{th} covariate is missing. (Formally, $\underset{\sim}{x}^*$ is an element of $\{R, *\}^p$.) Then $e^* = pr(z=1|\underset{\sim}{x}^*)$ is a generalized propensity score. The following theorem and corollary show the e^* has balancing properties that are similar to the balancing properties of the propensity score e.

Theorem 1. $$\underset{\sim}{x}^* \perp\!\!\!\perp z | e^*$$

Proof: Identical to the proof of theorem 1 of Rosenbaum and Rubin (1983a), with $\underset{\sim}{x}^*$ in place of $\underset{\sim}{x}$ and e^* in place of e.

Theorem 1 implies that subclassification on the generalized propensity score e^* balances the observed covariate information and the pattern of missing covariates. Note that Theorem 1 does not generally imply that subclassification on e^* balances the unobserved coordinates of $\underset{\sim}{x}$; that is, Theorem 1 does not generally imply

$$\underset{\sim}{x} \perp\!\!\!\perp z \mid e^* \ .$$

The consequences of Theorem 1 are clearest when there are only two patterns of missing data, with $\underset{\sim}{x} = (\underset{\sim}{x}_1, \underset{\sim}{x}_2)$, where $\underset{\sim}{x}_1$ is always observed and $\underset{\sim}{x}_2$ is sometimes missing. Let $c = 1$ when $\underset{\sim}{x}_2$ is observed, and let $c = 0$ when $\underset{\sim}{x}_2$ is missing. Then $e^* = pr(z=1|\underset{\sim}{x}_1, \underset{\sim}{x}_2, c=1)$ for units with $\underset{\sim}{x}_2$ observed, and $e^* = pr(z=1|\underset{\sim}{x}_1, c=0)$ for units with $\underset{\sim}{x}_2$ missing. Subclasses of units may be formed using e^*, ignoring the pattern of missing data.

Corollary 1.1.

A. For units with $\underset{\sim}{x}_2$ missing, there is balance on $\underset{\sim}{x}_1$ at each value of e^*, that is,

$$\underset{\sim}{x}_1 \perp\!\!\!\perp z | e^*, \ c = 0 \ .$$

B. For units with $\underset{\sim}{x}_2$ observed, there is balance on $(\underset{\sim}{x}_1, \underset{\sim}{x}_2)$ at each value of e^*, that is,

$$(\underset{\sim}{x}_1, \underset{\sim}{x}_2) \perp\!\!\!\perp z | e^*, \ c = 1 \ .$$

C. There is balance on $\underset{\sim}{x}_1$ at each value of e^*, that is,

$$\underset{\sim}{x}_1 \perp\!\!\!\perp z | e^* \ .$$

D. The frequency of missing data is balanced at each value of e^*, that is,

$$c \perp\!\!\!\perp z | e^* \ .$$

Proof: Parts A and B follow immediately from Theorem 1 of Rosenbaum and Rubin (1983a), and Parts C and D following immediately from Theorem 1 above.

In practice, we may estimate e^* in several ways. In a large study with only a few patterns of missing data, we may use a separate logit model for each pattern of missing data. In general, however, there are 2^p potential patterns of missing data with p covariates. If the covariates are discrete, then we may estimate e^* by treating the $*$ as an additional category for each of the p covariates.

Apendix B: The Effectiveness of Subclassification on the Propensity Score in Removing Bias

Cochran (1968) studied the effectiveness of univariate subclassification in removing bias in observational studies. In this appendix, we show how Cochran's results are related to subclassification on the propensity score.

Let $f = f(\underset{\sim}{x})$ be any scalar valued function of $\underset{\sim}{x}$. The initial bias in f is $B_I = E(f|z=1) - E(f|z=0)$. The bias in f after subclassification on the propensity score and direct adjustment with subclass total weights is

$$B_S = \sum_{j=1}^{J} \{E(f|z=1,\ e \in I_j) - E(f|z=0,\ e \in I_j)\}pr(e \in I_j)$$

where there are J subclasses, and I_j is the set of values of e that define the j^{th} subclass. The percent reduction in bias in f due to subclassification on the propensity score is $100[1 - \frac{B_S}{B_I}]$.

Cochran's (1968) results do not directly apply to subclassification on the propensity score since his work is concerned with the percent reduction in bias in f after subclassification on f, rather than the percent reduction in bias in f after subclassification on e. Nonetheless, as the following theorem shows, Cochran's results are applicable providing (a) the conditional expectation of f given e, that is $E(f|e) = \bar{f}$, say, is a monotone function of e, and (b) $\bar{f}$ has one of the distributions studied by Cochran. In particular, under these conditions, subclassification at the quintiles of the distribution of the propensity score, e, will produce approximately a 90% reduction in the bias of f. Note that in the following theorem, Cochran's (1968) results apply directly to the problem of determining the percent reduction in bias in $\bar{f}$ after subclassification on $\bar{f}$.

Theorem 2: The percent reduction in the bias, $100(1 - \frac{B_S}{B_I})$, in f following subclassification at specified quantiles of the distribution of the propensity score, e, equals the percent reduction in bias in $\bar{f}$ after subclassification at the same quantiles of the distribution of $\bar{f}$, providing $\bar{f}$ is a strictly monotone function of e.

Proof: First, we show that within a subclass defined by $e \in S$, the bias in f equals the bias in $\bar{f}$; that is, we show that

(B.1) $\quad E(f|e \in S, z=1) - E(f|e \in S, z=0) = E(\bar{f}|e \in S, z=1) - E(\bar{f}|e \in S, z=0)$.

To show this it is sufficient to observe that for $t = 0, 1$,

$$\begin{aligned} E(f|e \in S, z=t) &= E\{E(f|e, e \in S, z=t)|e \in S, z=t\} \\ &= E\{E(f|e)|e \in S, z=t\} \\ &= E(\bar{f}|e \in S, z=t) \end{aligned}$$

where the second equality follows from the fact that e is the propensity score (i.e., from equation (1)).

From (B.1) with $S = (-\infty,\infty)$, it follows that the initial bias in f equals the initial bias in $\bar{f}$. To complete the proof, we need to show that the bias in f after subclassification on e equals the bias in $\bar{f}$ after subclassification on $\bar{f}$. Since by assumption $\bar{f}$ is a strictly monotone function of e, subclasses defined at specified quantiles of the distribtuion of e contain exactly the same units as subclasses defined at the same quantiles of the distribution of $\bar{f}$. It follows from this observation and (B.1) that the bias in f within each subclass defined by e equals the bias in $\bar{f}$ within each subclass defined by $\bar{f}$. Since (a) the initial biases in f and $\bar{f}$ are equal, (b) the subclasses formed from e contain the same units as the subclasses formed from $\bar{f}$, and (c) within each subclass, the bias in f equals the bias in $\bar{f}$, it follows that the percent reduction in bias in f after subclassification on e equals the percent reduction in bias in $\bar{f}$ after subclassification on $\bar{f}$. //

Acknowledgement

The authors acknowledge valuable conversations with A. P. Dempster on the issues discussed in this paper.

REFERENCES

Benson, H. and McCallie, D. (1979). Angina pectoris and the placebo effect. New England Journal of Medicine, 300, 1424-1428.

Cochran, W. G. (1965). The planning of observational studies of human populations. Journal of the Royal Statistical Society, Series A 128, 234-255.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observatinal studies. Biometrics, 24, 205-213.

Cox, D. R. (1970). The Analysis of Binary Data. London: Methuen.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). Journal of the Royal Statistical Society, Series B, 41, 1-31.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data using the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-38.

Miettinen, O. (1976). Stratification by a multivariate confounder score. American Journal of Epidemiology, 104 : 609-620.

Mosteller, C. F. & Tukey, J. W. (1977). Data Analysis and Regression. Reading, MA: Addison-Wesley.

Rosenbaum, P. R. (1982). Testing the assumption of strongly ignorable treatment assignment in observational studies: A review within a general framework. Submitted to the Journal of the American Statistical Association.

Rosenbaum, P. R. & Rubin, D. B. (1983,a). The central role of the propensity score in observational studies for causal effects. To appear in Biometrika, 70, #1.

Rosenbaum, P. R. & Rubin, D. B. (1983,b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. To appear in the Journal of the Royal Statistical Society, Series B, 45, #2.

Rubin, D. B. (1976). Inference and missing data. Biometrika 63, 581-592.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. Annals of Statistics, 6, 34-58.

Tukey, J. W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley.

PRR/DBR/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2468	2. GOVT ACCESSION NO. AD-A127760	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Balanced Subclassification in Observational Studies Using the Propensity Score: A Case Study		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Paul R. Rosenbaum and Donald B. Rubin		8. CONTRACT OR GRANT NUMBER(s) ~~P3D-CA-14520~~ DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS See Item 18 below		12. REPORT DATE January 1983 13. NUMBER OF PAGES 23
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Cancer Institute
9000 Rockville Pike
Bethesda, MD 20205

Educational Testing Service
Carter Road
Princeton, NJ 08541

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Observational studies; bias reduction; stratification; logistic models; log linear models; direct adjustment; balancing score.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Previous theoretical arguments have shown that subclassification on the scalar propensity score will balance all observed covariates. The procedure is illustrated in a large observational study of treatments for coronary artery disease. Five subclasses are constructed that balance 74 covariates. Balanced subclassification is combined with model-based adjustments to provide estimates of treatment effects within subpopulations. Two appendices address theoretical issues:

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

(next page)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT (continued)

(A) propensity scores from incomplete data, and (B) the effectiveness of subclassification on the propensity score.

END

DATE
FILMED

5-83

DTIC